A slightly different version appeared in *Proceedings of the 37th ACL*, 1999

Measures of Distributional Similarity

Lillian Lee
Department of Computer Science
Cornell University
Ithaca, NY 14853-7501
llee@cs.cornell.edu

Abstract

We study distributional similarity measures for the purpose of improving probability estimation for unseen cooccurrences. Our contributions are three-fold: an empirical comparison of a broad range of measures; a classification of similarity functions based on the information that they incorporate; and the introduction of a novel function that is superior at evaluating potential proxy distributions.

1 Introduction

An inherent problem for statistical methods in natural language processing is that of sparse data — the inaccurate representation in any training corpus of the probability of low frequency events. In particular, reasonable events that happen to not occur in the training set may mistakenly be assigned a probability of zero. These *unseen* events generally make up a substantial portion of novel data; for example, Essen and Steinbiss (1992) report that 12% of the test-set bigrams in a 75%-25% split of one million words did not occur in the training partition.

We consider here the question of how to estimate the conditional cooccurrence probability $P(v|n)$ of an unseen word pair (n, v) drawn from some finite set $N \times V$. Two state-of-the-art technologies are Katz's (1987) *backoff* method and Jelinek and Mercer's (1980) interpolation method. Both use $P(v)$ to estimate $P(v|n)$ when (n, v) is unseen, essentially ignoring the identity of n.

An alternative approach is *distance-weighted averaging*, which arrives at an estimate for unseen cooccurrences by combining estimates for cooccurrences involving similar words:[1]

$$\hat{P}(v|n) = \frac{\sum_{m \in \mathcal{S}(n)} \mathrm{sim}(n,m) P(v|m)}{\sum_{m \in \mathcal{S}(n)} \mathrm{sim}(n,m)}, \qquad (1)$$

where $\mathcal{S}(n)$ is a set of candidate similar words and $\mathrm{sim}(n, m)$ is a function of the similarity between n and m. We focus on *distributional* rather than semantic similarity (e.g., Resnik (1995)) because the goal of distance-weighted averaging is to smooth probability distributions — although the words "chance" and "probability" are synonyms, the former may not be a good model for predicting what cooccurrences the latter is likely to participate in.

There are many plausible measures of distributional similarity. In previous work (Dagan et al., 1999), we compared the performance of three different functions: the Jensen-Shannon divergence (total divergence to the average), the L_1 norm, and the confusion probability. Our experiments on a frequency-controlled pseudoword disambiguation task showed that using any of the three in a distance-weighted averaging scheme yielded large improvements over Katz's backoff smoothing method in predicting unseen coocurrences. Furthermore, by using a restricted version of model (1) that stripped incomparable parameters, we were able to empirically demonstrate that the confusion probability is fundamentally worse at selecting useful similar words. D. Lin also found that the choice of similarity function can affect the quality of automatically-constructed thesauri to a statistically significant degree (1998a) and the ability to determine common morphological roots by as much as 49% in precision (1998b).

[1] The term "similarity-based", which we have used previously, has been applied to describe other models as well (L. Lee, 1997; Karov and Edelman, 1998).

These empirical results indicate that investigating different similarity measures can lead to improved natural language processing. On the other hand, while there have been many similarity measures proposed and analyzed in the information retrieval literature (Jones and Furnas, 1987), there has been some doubt expressed in that community that the choice of similarity metric has any practical impact:

> Several authors have pointed out that the difference in retrieval performance achieved by different measures of association is insignificant, providing that these are appropriately normalised. (van Rijsbergen, 1979, pg. 38)

But no contradiction arises because, as van Rijsbergen continues, "one would expect this since most measures incorporate the same information". In the language-modeling domain, there is currently no agreed-upon best similarity metric because there is no agreement on what the "same information" — the key data that a similarity function should incorporate — is.

The overall goal of the work described here was to discover these key characteristics. To this end, we first compared a number of common similarity measures, evaluating them in a parameter-free way on a decision task. When grouped by average performance, they fell into several coherent classes, which corresponded to the extent to which the functions focused on the intersection of the *supports* (regions of positive probability) of the distributions. Using this insight, we developed an information-theoretic metric, the *skew divergence*, which incorporates the support-intersection data in an asymmetric fashion. This function yielded the best performance overall: an average error rate reduction of 4% (significant at the .01 level) with respect to the Jensen-Shannon divergence, the best predictor of unseen events in our earlier experiments (Dagan et al., 1999).

Our contributions are thus three-fold: an empirical comparison of a broad range of similarity metrics using an evaluation methodology that factors out inessential degrees of freedom; a proposal, building on this comparison, of a characteristic for classifying similarity functions; and the introduction of a new similarity metric incorporating this characteristic that is superior at evaluating potential proxy distributions.

2 Distributional Similarity Functions

In this section, we describe the seven distributional similarity functions we initally evaluated.[2] For concreteness, we choose N and V to be the set of nouns and the set of transitive verbs, respectively; a cooccurrence pair (n, v) results when n appears as the head noun of the direct object of v. We use P to denote probabilities assigned by a base language model (in our experiments, we simply used unsmoothed relative frequencies derived from training corpus counts).

Let n and m be two nouns whose distributional similarity is to be determined; for notational simplicity, we write $q(v)$ for $P(v|n)$ and $r(v)$ for $P(v|m)$, their respective conditional verb cooccurrence probabilities.

Figure 1 lists several familiar functions. The cosine metric and Jaccard's coefficient are commonly used in information retrieval as measures of association (Salton and McGill, 1983). Note that Jaccard's coefficient differs from all the other measures we consider in that it is essentially *combinatorial*, being based only on the sizes of the supports of q, r, and $q \cdot r$ rather than the actual values of the distributions.

Previously, we found the *Jensen-Shannon divergence* (Rao, 1982; J. Lin, 1991) to be a useful measure of the distance between distributions:

$$JS(q,r) = \frac{1}{2}\left[D\left(q \,\middle\|\, \mathrm{avg}_{q,r}\right) + D\left(r \,\middle\|\, \mathrm{avg}_{q,r}\right)\right] .$$

The function D is the *KL divergence*, which measures the (always nonnegative) average inefficiency in using one distribution to code for another (Cover and Thomas, 1991):

$$D(p_1(V) \,\|\, p_2(V)) = \sum_v p_1(v) \log \frac{p_1(v)}{p_2(v)} .$$

The function $\mathrm{avg}_{q,r}$ denotes the average distribution $\mathrm{avg}_{q,r}(v) = (q(v) + r(v))/2$; observe that its use ensures that the Jensen-Shannon divergence is always defined. In contrast, $D(q||r)$ is undefined if q is not absolutely continuous with respect to r (i.e., the support of q is not a subset of the support of r).

[2]Strictly speaking, some of these functions are dissimilarity measures, but each such function f can be recast as a similarity function via the simple transformation $C - f$, where C is an appropriate constant. Whether we mean f or $C - f$ should be clear from context.

$$
\begin{array}{rrcl}
\text{Euclidean distance} & L_2(q,r) & = & \sqrt{\sum_v (q(v)-r(v))^2} \\
L_1 \text{ norm} & L_1(q,r) & = & \sum_v |q(v)-r(v)| \\
\text{cosine} & \cos(q,r) & = & \frac{\sum_v q(v)r(v)}{\sqrt{\sum_v q(v)^2}\sqrt{\sum_v r(v)^2}} \\
\text{Jaccard's coefficient} & \mathrm{Jac}(q,r) & = & \frac{|\{v \,:\, q(v)>0 \text{ and } r(v)>0\}|}{|\{v \mid q(v)>0 \text{ or } r(v)>0\}|}
\end{array}
$$

Figure 1: Well-known functions

The *confusion probability* has been used by several authors to smooth word cooccurrence probabilities (Sugawara et al., 1985; Essen and Steinbiss, 1992; Grishman and Sterling, 1993); it measures the degree to which word m can be substituted into the contexts in which n appears. If the base language model probabilities obey certain Bayesian consistency conditions (Dagan et al., 1999), as is the case for relative frequencies, then we may write the confusion probability as follows:

$$\mathrm{conf}(q,r,P(m)) = \sum_v q(v)r(v)\frac{P(m)}{P(v)} .$$

Note that it incorporates unigram probabilities as well as the two distributions q and r.

Finally, *Kendall's* τ, which appears in work on clustering similar adjectives (Hatzivassiloglou and McKeown, 1993; Hatzivassiloglou, 1996), is a nonparametric measure of the association between random variables (Gibbons, 1993). In our context, it looks for correlation between the behavior of q and r on pairs of verbs. Three versions exist; we use the simplest, τ_a, here:

$$\tau(q,r) = \sum_{v_1,v_2} \frac{\mathrm{sign}\left[(q(v_1)-q(v_2))(r(v_1)-r(v_2))\right]}{2\binom{|V|}{2}},$$

where $\mathrm{sign}(x)$ is 1 for positive arguments, -1 for negative arguments, and 0 at 0. The intuition behind Kendall's τ is as follows. Assume all verbs have distinct conditional probabilities. If sorting the verbs by the likelihoods assigned by q yields exactly the same ordering as that which results from ranking them according to r, then $\tau(q,r)=1$; if it yields exactly the opposite ordering, then $\tau(q,r)=-1$. We treat a value of -1 as indicating extreme dissimilarity.[3]

It is worth noting at this point that there are several well-known measures from the NLP literature that we have omitted from our experiments. Arguably the most widely used is the *mutual information* (Hindle, 1990; Church and Hanks, 1990; Dagan et al., 1995; Luk, 1995; D. Lin, 1998a). It does not apply in the present setting because it does not measure the similarity between two arbitrary probability distributions (in our case, $P(V|n)$ and $P(V|m)$), but rather the similarity between a joint distribution $P(X_1,X_2)$ and the corresponding product distribution $P(X_1)P(X_2)$. Hamming-type metrics (Cardie, 1993; Zavrel and Daelemans, 1997) are intended for data with symbolic features, since they count feature label mismatches, whereas we are dealing feature values that are probabilities. Variations of the *value difference metric* (Stanfill and Waltz, 1986) have been employed for supervised disambiguation (Ng and H.B. Lee, 1996; Ng, 1997); but it is not reasonable in language modeling to expect training data tagged with correct probabilities. The *Dice coefficient* (Smadja et al., 1996; D. Lin, 1998a, 1998b) (Kay and Röscheisen, 1993) is monotonic in Jaccard's coefficient (van Rijsbergen, 1979), so its inclusion in our experiments would be redundant. Finally, we did not use the KL divergence because it requires a smoothed base language model.

[3]Zero would also be a reasonable choice, since it indicates zero correlation between q and r. However, it would then not be clear how to average in the estimates of negatively correlated words in equation (1).

3 Empirical Comparison

We evaluated the similarity functions introduced in the previous section on a binary decision task, using the same experimental framework as in our previous preliminary comparison (Dagan et al., 1999). That is, the data consisted of the verb-object cooccurrence pairs in the 1988 Associated Press newswire involving the 1000 most frequent nouns, extracted via Church's (1988) and Yarowsky's processing tools. 587,833 (80%) of the pairs served as a training set from which to calculate base probabilities. From the other 20%, we prepared test sets as follows: after discarding pairs occurring in the training data (after all, the point of similarity-based estimation is to deal with unseen pairs), we split the remaining pairs into five partitions, and replaced each noun-verb pair (n, v_1) with a noun-verb-verb triple (n, v_1, v_2) such that $P(v_2) \approx P(v_1)$. The task for the language model under evaluation was to reconstruct which of (n, v_1) and (n, v_2) was the original cooccurrence. Note that by construction, (n, v_1) was always the correct answer, and furthermore, methods relying solely on unigram frequencies would perform no better than chance. Test-set performance was measured by the error rate, defined as

$$\frac{1}{T}(\#\text{ of incorrect choices } + (\#\text{ of ties})/2)\,,$$

where T is the number of test triple tokens in the set, and a tie results when both alternatives are deemed equally likely by the language model in question.

To perform the evaluation, we incorporated each similarity function into a simple decision rule as follows. As above, let (n, v_1, v_2) be a test instance. For a given similarity measure f and neighborhood size k, let $\mathcal{S}_{f,k}(n)$ denote the k most similar words to n according to f. We define the *evidence* $E_{f,k}(n, v_1)$ for v_1 as the number of neighbors $m \in \mathcal{S}_{f,k}(n)$ such that $P(v_1|m) > P(v_2|m)$; similarly, the evidence for v_2 is the number of the k closest neighbors that favor v_2 over v_1. Then, the decision rule is to choose the verb alternative with the greatest evidence.

The reason we used a restricted version of the distance-weighted averaging model was that we sought to discover fundamental differences in behavior. Because we have a binary decision task, $E_{f,k}(n, v_1)$ simply counts the number of k nearest neighbors to n that make the right decision. If we have two functions f and g such that $E_{f,k}(n, v_1) > E_{g,k}(n, v_1)$, then the k most similar words according to f are on the whole better predictors than the k most similar words according to g; hence, f induces an inherently better similarity ranking for distance-weighted averaging. The difficulty with using the full model (Equation (1)) for comparison purposes is that fundamental differences can be obscured by issues of weighting. For example, suppose the probability estimate $\sum_v (2 - L_1(q, r)) \cdot r(v)$ (suitably normalized) performed poorly. We would not be able to tell whether the cause was an inherent deficiency in the L_1 norm or just a poor choice of weight function — perhaps $(2 - L_1(q, r))^2$ would have yielded better estimates.

Figure 2 shows how the average error rate varies with k for the seven similarity metrics introduced above. As previously mentioned, a steeper slope indicates a better similarity ranking.

All the curves have a generally upward trend but always lie far below backoff (51% error rate). They meet at $k = 1000$ because $\mathcal{S}_{f,1000}(n)$ is always the set of all nouns. We see that the functions fall into four groups: (1) the L_2 norm; (2) Kendall's τ; (3) the confusion probability and the cosine metric; and (4) the L_1 norm, Jensen-Shannon divergence, and Jaccard's coefficient.

We can account for the similar performance of various metrics by analyzing how they incorporate information from the intersection of the supports of q and r. (Recall that we are using q and r for the conditional verb cooccurrrence probabilities of two nouns n and m.) Consider the following supports (illustrated in Figure 3):

$$\begin{aligned} V_q &= \{v \in V : q(v) > 0\} \\ V_r &= \{v \in V : r(v) > 0\} \\ V_{qr} &= \{v \in V : q(v)r(v) > 0\} = V_q \cap V_r \end{aligned}$$

We can rewrite the similarity functions from Section 2 in terms of these sets, making use of the identities $\sum_{v \in V_q \setminus V_{qr}} q(v) + \sum_{v \in V_{qr}} q(v) = \sum_{v \in V_r \setminus V_{qr}} r(v) + \sum_{v \in V_{qr}} r(v) = 1$. Table 1 lists these alternative forms in order of performance.

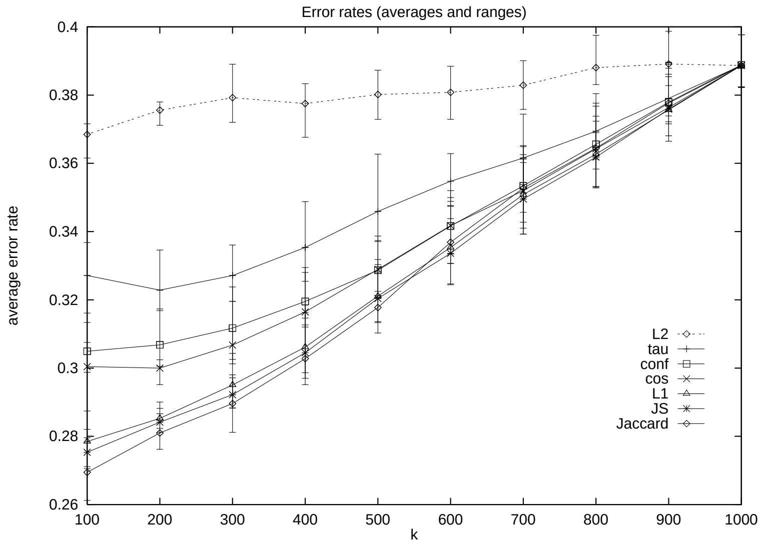

Figure 2: Similarity metric performance. Errorbars denote the range of error rates over the five test sets. Backoff's average error rate was 51%.

$$L_2(q,r) = \sqrt{\sum_{V_q} q(v)^2 - 2\sum_{V_{qr}} q(v)r(v) + \sum_{V_r} r(v)^2}$$

$$\tau(q,r)\cdot 2\binom{|V|}{2} = 2\,|V_{qr}|\,|V \setminus (V_q \cup V_r)| - 2\,|V_q \setminus V_{qr}|\,|V_r \setminus V_{qr}| + \sum_{v_1 \in (V_q \triangle V_r)} \sum_{v_2 \in V_{qr}} \operatorname{sign}[(q(v_1)-q(v_2))(r(v_1)-r(v_2))] + \sum_{v_1 \in V_{qr}} \sum_{v_2 \in V_q \cup V_r} \operatorname{sign}[(q(v_1)-q(v_2))(r(v_1)-r(v_2))]$$

$$\mathrm{conf}(q,r,P(m)) = P(m) \sum_{v \in V_{qr}} q(v)r(v)/P(v)$$

$$\cos(q,r) = \sum_{v \in V_{qr}} q(v)r(v)\left(\sum_{v \in V_q} q(v)^2 \sum_{v \in V_r} r(v)^2\right)^{-1/2}$$

$$L_1(q,r) = 2 - \sum_{v \in V_{qr}} \left(|q(v)-r(v)| - q(v) - r(v)\right)$$

$$JS(q,r) = \log 2 + \frac{1}{2} \sum_{v \in V_{qr}} \left(h(q(v)+r(v)) - h(q(v)) - h(r(v))\right), \qquad h(x) = -x\log x$$

$$\mathrm{Jac}(q,r) = |V_{qr}| \,/\, |V_q \cup V_r|$$

Table 1: Similarity functions, written in terms of sums over supports and grouped by average performance. $\setminus$ denotes set difference; $\triangle$ denotes symmetric set difference.

We see that for the non-combinatorial functions, the groups correspond to the degree to which the measures rely on the verbs in V_{qr}. The Jensen-Shannon divergence and the L_1 norm can be computed simply by knowing the values of q and r on V_{qr}. For the cosine and the confusion probability, the distribution values on V_{qr} are key, but other information is also incorporated. The statistic τ_a takes into account all verbs, including those that occur neither with n nor m. Finally, the Euclidean distance is quadratic in verbs outside V_{qr}; indeed, Kaufman

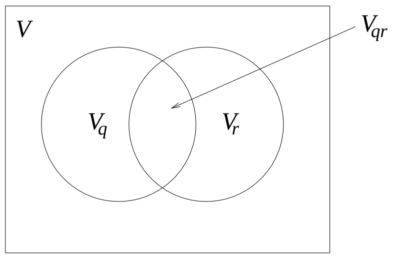

Figure 3: Supports on V

and Rousseeuw (1990) note that it is "extremely sensitive to the effect of one or more outliers" (pg. 117).

The superior performance of $\mathrm{Jac}(q, r)$ seems to underscore the importance of the set V_{qr}. Jaccard's coefficient ignores the values of q and r on V_{qr}; but we see that simply knowing the size of V_{qr} relative to the supports of q and r leads to good rankings.

4 The Skew Divergence

Based on the results just described, it appears that it is desirable to have a similarity function that focuses on the verbs that cooccur with both of the nouns being compared. However, we can make a further observation: with the exception of the confusion probability, all the functions we compared are symmetric, that is, $f(q, r) = f(r, q)$. But the substitutability of one word for another need not symmetric. For instance, "fruit" may be the best possible approximation to "apple", but the distribution of "apple" may not be a suitable proxy for the distribution of "fruit".[4]

In accordance with this insight, we developed a novel asymmetric generalization of the KL divergence, the *α-skew divergence*:

$$s_\alpha(q, r) = D(r \parallel \alpha \cdot q + (1 - \alpha) \cdot r)$$

for $0 \le \alpha \le 1$. It can easily be shown that s_α depends only on the verbs in V_{qr}. Note that at $\alpha = 1$, the skew divergence is exactly the KL divergence, and $s_{1/2}$ is twice one of the summands of JS (note that it is still asymmetric).

We can think of α as a degree of confidence in the empirical distribution q; or, equivalently, $(1 - \alpha)$ can be thought of as controlling the amount by which one smooths q by r. Thus, we can view the skew divergence as an approximation to the KL divergence to be used when sparse data problems would cause the latter measure to be undefined.

Figure 4 shows the performance of s_α for $\alpha = .99$. It performs better than all the other functions; the difference with respect to Jaccard's coefficient is statistically significant, according to the paired t-test, at all k (except $k = 1000$), with significance level .01 at all k except 100, 400, and 1000.

5 Discussion

In this paper, we empirically evaluated a number of distributional similarity measures, including the skew divergence, and analyzed their information sources. We observed that the ability of a similarity function $f(q, r)$ to select useful nearest neighbors appears to be correlated with its focus on the intersection V_{qr} of the supports of q and r. This is of interest from a computational point of view because V_{qr} tends to be a relatively small subset of V, the set of all verbs. Furthermore, it suggests downplaying the role of negative information, which is encoded by verbs appearing with exactly one noun, although the Jaccard coefficient does take this type of information into account.

Our explicit division of V-space into various support regions has been implicitly considered in other work. Smadja et al. (1996) observe that for two potential mutual translations X and Y, the fact that X occurs with translation Y indicates association; X's occurring with a translation other than Y decreases one's belief in their association; but the absence of both X and Y yields no information. In essence, Smadja et al. argue that information from the union of supports, rather than the just the intersection, is important. D. Lin (1997; 1998a) takes an axiomatic approach to determining the characteristics of a good similarity measure. Starting with a formalization (based on certain assumptions) of the intuition that the similarity between two events depends on both their commonality and their differences, he derives a unique similarity function schema. The definition of commonality is left to the user (several different definitions are proposed for differ-

[4] On a related note, an anonymous reviewer cited the following example from the psychology literature: we can say Smith's lecture is like a sleeping pill, but "not the other way round".

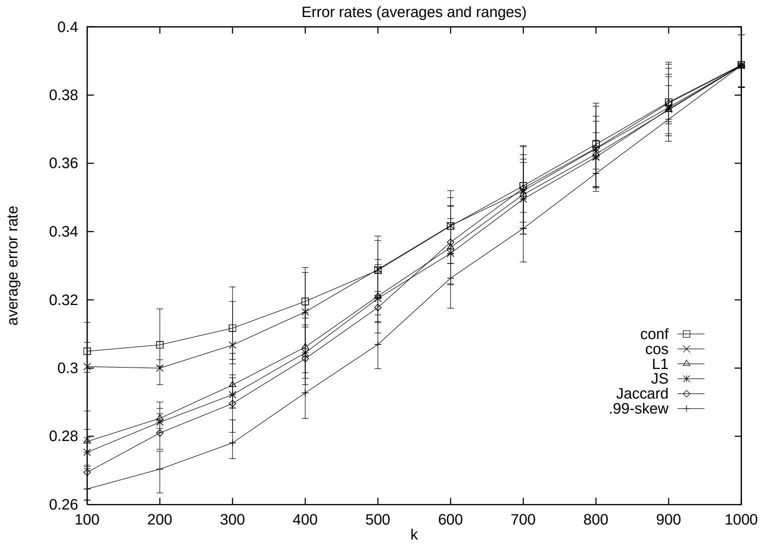

Figure 4: Performance of the skew divergence with respect to the best functions from Figure 2.

ent tasks).

We view the empirical approach taken in this paper as complementary to Lin's. That is, we are working in the context of a particular application, and, while we have no mathematical certainty of the importance of the "common support" information, we did not assume it *a priori*; rather, we let the performance data guide our thinking.

Finally, we observe that the skew metric seems quite promising. We conjecture that appropriate values for α may inversely correspond to the degree of sparseness in the data, and intend in the future to test this conjecture on larger-scale prediction tasks. We also plan to evaluate skewed versions of the Jensen-Shannon divergence proposed by Rao (1982) and J. Lin (1991).

6 Acknowledgements

Thanks to Claire Cardie, Jon Kleinberg, David McAllester, Fernando Pereira, and Stuart Shieber for helpful discussions, the anonymous reviewers for their insightful comments, Fernando Pereira for access to computational resources at AT&T, and Stuart Shieber for the opportunity to pursue this work at Harvard University. This material is based upon work supported by the National Science Foundation under Grant No. IRI9712068.

References

Claire Cardie. 1993. A case-based approach to knowledge acquisition for domain-specific sentence analysis. In *11th National Conference on Artifical Intelligence*, pages 798–803.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Kenneth W. Church. 1988. A stochastic parts program and noun phrase parser for unrestricted text. In *Second Conference on Applied Natural Language Processing*, pages 136–143.

Thomas M. Cover and Joy A. Thomas. 1991. *Elements of Information Theory*. John Wiley.

Ido Dagan, Shaul Marcus, and Shaul Markovitch. 1995. Contextual word similarity and estimation from sparse data. *Computer Speech and Language*, 9:123–152.

Ido Dagan, Lillian Lee, and Fernando Pereira. 1999. Similarity-based models of cooccurrence probabilities. *Machine Learning*, 34(1-3):43–69.

Ute Essen and Volker Steinbiss. 1992. Co-occurrence smoothing for stochastic language modeling. In *ICASSP 92*, volume 1, pages 161–164.

Jean Dickinson Gibbons. 1993. *Nonparametric Measures of Association*. Number 07-091 in Sage University Paper series on Quantitative Applications in the Social Sciences. Sage, Newberry Park, CA.

Ralph Grishman and John Sterling. 1993. Smoothing of automatically generated selectional constraints. In *Human Language Technology: Proceedings of the ARPA Workshop*, pages 254–259.

Vasileios Hatzivassiloglou and Kathleen McKeown. 1993. Towards the automatic identification of adjectival scales: Clustering of adjectives according to meaning. In *31st Annual Meeting of the ACL*, pages 172–182.

Vasileios Hatzivassiloglou. 1996. Do we need linguistics when we have statistics? A comparative analysis of the contributions of linguistic cues to a statistical word grouping system. In Judith L. Klavans and Philip Resnik, editors, *The Balancing Act*, pages 67–94. MIT Press.

Don Hindle. 1990. Noun classification from predicate-argument structures. In *28th Annual Meeting of the ACL*, pages 268–275.

Frederick Jelinek and Robert L. Mercer. 1980. Interpolated estimation of Markov source parameters from sparse data. In *Proceedings of the Workshop on Pattern Recognition in Practice*.

William P. Jones and George W. Furnas. 1987. Pictures of relevance. *Journal of the American Society for Information Science*, 38(6):420–442.

Yael Karov and Shimon Edelman. 1998. Similarity-based word sense disambiguation. *Computational Linguistics*, 24(1):41–59.

Slava M. Katz. 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-35(3):400–401, March.

Leonard Kaufman and Peter J. Rousseeuw. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley and Sons.

Martin Kay and Martin Röscheisen. 1993. Text-translation alignment. *Computational Linguistics*, 19(1):121–142.

Lillian Lee. 1997. *Similarity-Based Approaches to Natural Language Processing*. Ph.D. thesis, Harvard University.

Jianhua Lin. 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.

Dekang Lin. 1997. Using syntactic dependency as local context to resolve word sense ambiguity. In *35th Annual Meeting of the ACL*, pages 64–71.

Dekang Lin. 1998a. Automatic retrieval and clustering of similar words. In *COLING-ACL '98*, pages 768–773.

Dekang Lin. 1998b. An information theoretic definition of similarity. In *Machine Learning: Proceedings of the Fiftheenth International Conference (ICML '98)*.

Alpha K. Luk. 1995. Statistical sense disambiguation with relatively small corpora using dictionary definitions. In *33rd Annual Meeting of the ACL*, pages 181–188.

Hwee Tou Ng and Hian Beng Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *34th Annual Meeting of the ACL*, pages 40–47.

Hwee Tou Ng. 1997. Exemplar-based word sense disambiguation: Some recent improvements. In *Second Conference on Empirical Methods in Natural Language Processing (EMNLP-2)*, pages 208–213.

C. Radhakrishna Rao. 1982. Diversity: Its measurement, decomposition, apportionment and analysis. *Sankyhā: The Indian Journal of Statistics*, 44(A):1–22.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of IJCAI-95*, pages 448–453.

Gerard Salton and Michael J. McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill.

Frank Smadja, Kathleen R. McKeown, and Vasileios Hatzivassiloglou. 1996. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, 22(1):1–38.

Craig Stanfill and David Waltz. 1986. Toward memory-based reasoning. *Communications of the ACM*, 29(12):1213–1228.

K. Sugawara, M. Nishimura, K. Toshioka, M. Okochi, and T. Kaneko. 1985. Isolated word recognition using hidden Markov models. In *ICASSP 85*, pages 1–4.

C. J. van Rijsbergen. 1979. *Information Retrieval*. Butterworths, second edition.

Jakub Zavrel and Walter Daelemans. 1997. Memory-based learning: Using similarity for smoothing. In *35th Annual Meeting of the ACL*, pages 436–443.